5 Minute Summary

Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types: agglomerative and divisive.

Agglomerative Clustering

Agglomerative clustering is a "bottom-up" approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. The process begins with �n clusters and iteratively merges them until one cluster (or �k clusters, if stopping early) is left.

Divisive Clustering

Divisive clustering is a "top-down" approach starting with all observations in a single cluster. A cluster is divided into smaller clusters, which are recursively split until each observation ends up in its own cluster or until a stopping criterion is met.

Proximity Matrix

The proximity matrix stores the distances (similarities) between each pair of observations in the dataset. In agglomerative clustering, this matrix is updated at each iteration as the distances between new, merged clusters are calculated.

Dissimilarity Matrix

The dissimilarity matrix, similar to the proximity matrix, stores the dissimilarities between each pair of observations. It is used to determine how different two clusters or observations are, with higher values indicating greater dissimilarity.

Linkage Criteria

Linkage criteria determine the distance between sets of observations as a function of the pairwise distances between observations. Common linkage criteria include:

- **Single Linkage**: The minimum of all pairwise distances between two clusters.
- **Complete Linkage**: The maximum of all pairwise distances.
- **Average Linkage**: The average of all pairwise distances.
- **Ward's Linkage**: The increase in variance for the cluster being merged.

Reading a Dendrogram

A dendrogram is a tree-like diagram that records the sequences of merges or splits. The branches represent clusters that come together or split apart, with the length of the branches representing the distance (or dissimilarity) between clusters. To determine the number of clusters, one can cut the dendrogram at a desired level of dissimilarity, where the number of vertical lines intersected by the cut represents the number of clusters.

Deciding the Number of Clusters

The number of clusters in hierarchical clustering can be decided by inspecting the dendrogram and identifying a level where the cluster merge/split makes intuitive sense, or by using criteria such as the inconsistency coefficient, the silhouette score, or the elbow method.

Limitations of Agglomerative Clustering

- Computational complexity can be prohibitive for large datasets.
- The method is sensitive to noise and outliers.
- Once clusters are merged, they cannot be undone, which may lead to suboptimal solutions.
- The choice of linkage criteria can significantly affect the results, and there may not be a clear best choice for all datasets.

Additional Topics

- **Scalability Improvements**: Techniques like "mini-batch" can be applied to hierarchical clustering to improve scalability.

- **Evaluation Metrics**: Metrics like the silhouette coefficient, Calinski-Harabasz index, and Davies-Bouldin index can help evaluate the quality of clusters.

This summary encapsulates the key technical aspects of hierarchical clustering and related concepts, suitable for an audience already familiar with the basics.